PAC-Bayesian Stochastic Model Selection

David A. McAllester
AT&T Shannon Labs
180 Park Avenue
Florham Park, NJ 07932-0971
(973)360-8318
fax: (973)360-8970
email: dmac@research.att.com

June 26, 2001

Abstract

PAC-Bayesian learning methods combine the informative priors of Bayesian methods with distribution-free PAC guarantees. Stochastic model selection predicts a class label by stochastically sampling a classifier according to a "posterior distribution" on classifiers. This paper gives a PAC-Bayesian performance guarantee for stochastic model selection that is superior to analogous guarantees for deterministic model selection. The guarantee is stated in terms of the training error of the stochastic classifier and the KL-divergence of the posterior from the prior. It is shown that the posterior optimizing the performance guarantee is a Gibbs distribution. Simpler posterior distributions are also derived that have nearly optimal performance guarantees.

Keywords: PAC Learning, Model Averaging, Posterior Distribution, Gibbs Distribution, PAC-Baysian Learning.

1 INTRODUCTION

A PAC-Bayesian approach to machine learning attempts to combine the advantages of both PAC and Bayesian approaches [20, 15]. The Bayesian approach has the advantage of using arbitrary domain knowledge in the form of a Bayesian prior. The PAC approach has the advantage that one can prove guarantees for generalization error without assuming the truth of the prior. A PAC-Bayesian approach bases the bias of the learning algorithm on an arbitrary prior distribution, thus allowing the incorporation of domain knowledge, and yet provides a guarantee on generalization error that is independent of any truth of the prior.

PAC-Bayesian approaches are related to structural risk minimization (SRM) [11]. Here we interpret this broadly as describing any learning algorithm optimizing a tradeoff between the "complexity", "structure", or "prior probability" of the concept or model and the "goodness of fit", "description length", or "likelihood" of the training data. Under this interpretation of SRM, Bayesian algorithms that select a concept of maximum posterior probability (MAP algorithms) are viewed as a kind of SRM algorithm. Various approaches to SRM are compared both theoretically and experimentally by Kearns et al. in [11]. They give experimental evidence that Bayesian and MDL algorithms tend to over fit in experimental settings where the Bayesian assumptions fail. A PAC-Bayesian approach uses a prior distribution analogous to that used in MAP or MDL but provides a theoretical guarantee against over fitting independent of the truth of the prior.

Perhaps the simplest example of a PAC-Bayesian theorem is noted in [15]. Consider a countable class of concepts $f_1, f_2, f_3, \ldots$ where each concept f_i is a mapping from a set X to the two-valued set $\{0, 1\}$. Let P be an arbitrary "prior" probability distribution on these functions. Let D be any probability distribution on pairs $\langle x, y\rangle$ with $x \in X$ and $y \in \{0, 1\}$. We do not assume any relation between P and D. Define $\epsilon(f_i)$ to be the error rate of f_i, i.e., the probability over selecting $\langle x, y\rangle$ according to D that $f_i(x) \neq y$. Let S be a sample of m pairs drawn independently according to D and define $\hat{\epsilon}(f_i)$ to be the fraction of pairs $\langle x, y\rangle$ in S for which $f_i(x) \neq y$. Here $\hat{\epsilon}(f_i)$ is a measure of how well f_i fits the training data and $\log \frac{1}{P(f_i)}$ can be viewed as the "description length" of the concept f_i. It is noted in [15] that a simple combination of Chernoff and union bounds yields that with probability at least $1 - \delta$ over the choice of the sample S we have the following for all f_i.

$$\epsilon(f_i) \leq \hat{\epsilon}(f_i) + \sqrt{\frac{\ln \frac{1}{P(f_i)} + \ln \frac{1}{\delta}}{2m}} \tag{1}$$

This inequality justifies a concept selection algorithm which selects f^* to be the f_i minimizing the description-length vs. goodness-of-fit tradeoff in the right hand side. If there happens to be a low-description-length concept that fits well, the algorithm will perform well. If, however, all simple concepts fit poorly, the performance guarantee is poor. So in practice the probabilities

$P(f_i)$ should be arranged so that concepts which are a-priori viewed as likely to fit well are given high probability. Domain specific knowledge can be used in selecting the distribution P. This is precisely the sense in which P is analogous to a Bayesian prior — a concept f_i that is likely to fit well should be given high "prior probability" $P(f_i)$. Note, however, that the inequality (1) holds independent of any assumption about the relation between the distributions P and D.

Formula (1) is for model selection — algorithms that select a single model or concept. However, model selection is inferior to model averaging in certain applications. For example, in statistical language modeling for speech recognition one "smoothes" a trigram model with a bigram model and smoothes the bigram model with a unigram model. This smoothing is essential for minimizing the cross entropy between, say, the model and a test corpus of newspaper sentences. It turns out that smoothing in statistical language modeling is more naturally formulated as model averaging than as model selection. A smoothed language model is very large — it contains a full trigram model, a full bigram model and a full unigram model as parts. If one uses MDL to select the structure of a language model, selecting model parameters with maximum likelihood, the resulting structure is much smaller than that of a smoothed trigram model. Furthermore, the MDL model performs quite badly. A smoothed trigram model can be theoretically derived as a compact representation of a Bayesian mixture of an exponential number of (smaller) suffix tree models [18].

Model averaging can also be applied to decision trees that produce probabilities at their leaves rather than hard classifications. A common method of constructing decision trees is to first build an overly large tree which over fits the training data and then prune the tree in some way so as to get a smaller tree that does not over fit the data [19, 10]. For trees with probabilities at leaves, an alternative is to construct a weighted mixture of the subtrees of the original over fit tree. It is possible to construct a concise representation of a weighting over exponentially many different subtrees [3, 17, 9].

This paper is about stochastic model selection — algorithms that stochastically select a model according to a "posterior distribution" on the models. Stochastic model selection seems intermediate between model selection and model averaging — like model averaging it is based on a posterior distribution over models but it uses that distribution differently. Model averaging deterministically picks the value favored by a majority of models as weighted by the posterior. Stochastic model selection stochastically picks a single model according to the posterior distribution. The first main result of this paper is a bound on the performance of stochastic model selection that improves on (1) — stochastic model selection can be given better guarantees than deterministic model selection. Intuitively, model averaging should perform even better than stochastic model selection. But proving a PAC guarantee for model averaging superior to the PAC guarantees given here for stochastic model selection remains an open problem.

This paper also investigates the nature of the posterior distribution providing the best performance guarantee for stochastic model selection. It is shown that the optimal posterior is a Gibbs distribution. However, it is also shown that simpler posterior distributions are nearly optimal. Section 2 gives statements of the main results of this paper. Section 3 relates these results to previous work. The remaining sections present proofs.

2 Summary of the Main Results

Formula (1) applies to a countable class of concepts. It turns out that the guarantees on stochastic model selection hold for continuous classes as well, e.g., concepts with real-valued parameters. Here we assume a prior probability measure P on a possibly uncountable (continuous) concept class $\mathcal{C}$ and a sampling distribution D on a possibly uncountable set of instances $\mathcal{X}$. We also assume a measurable loss function l such that for any concept c and instance x we have $l(c,\ x) \in [0,\ 1]$. For example, we might have that concepts are predicates on instances and there is a target concept c_t such that $l(c,\ x)$ is 0 if $c(x) = c_t(x)$ and 1 otherwise. We define $l(c)$ to be the expectation over sampling an instance x of $l(c, x)$, i.e., $E_{x\sim D}l(c,\ x)$. We let S range over samples of m instances each drawn independently according to distribution D. We define $\hat{l}(c,\ S)$ to be $\frac{1}{m}\sum_{x\in S} l(c,\ x)$. If Q is a probability measure on concepts then $l(Q)$ denotes $E_{c\sim Q}l(c)$ and $\hat{l}(Q,\ S)$ denotes $E_{c\sim Q}\hat{l}(c_i,\ S)$. The notation $\forall^\delta S\ \Phi(S)$ signifies that the probability over the generation of the sample S of $\Phi(S)$ is at least $1-\delta$. For countable concept classes formula (1) generalizes as follows to any loss function l with $l(c,x) \in [0,1]$.

Lemma 1 (McAllester98) *For any probability distribution P on a countable rule class $\mathcal{C}$ we have the following.*

$$\forall^\delta S\ \ \forall c \in C\ \ l(c) \leq \hat{l}(c,\ S) + \sqrt{\frac{\ln \frac{1}{P(c)} + \ln \frac{1}{\delta}}{2m}}$$

As discussed in the introduction, this leads to a learning algorithm that selects the concept c^* minimizing the SRM tradeoff in the right hand side of the inequality. The first main result of this paper is a generalization of (1) to a uniform statement over *distributions* on an arbitrary concept class. The new bound involves the Kullback-Leibler divergence, denoted $D(Q||P)$, from distribution Q to distribution P. The quantity $D(Q||P)$ is defined to be $E_{c \sim Q} \ln \frac{dQ(c)}{dP(c)}$. The following is the first main result of this paper and is proved in section 4.

Theorem 1 *For any probability distribution (measure) on a possibly uncountable set $\mathcal{C}$ and any measurable loss function l we have the following where Q ranges over all distributions (measures) on $\mathcal{C}$.*

$$\forall^\delta S \ \ \forall Q \ \ l(Q) \leq \hat{l}(Q,\, S) + \sqrt{\frac{D(Q||P) + \ln \frac{1}{\delta} + \ln m + 2}{2m-1}}$$

Note that the definition of $l(Q)$, namely $E_{c \sim Q} l(c)$, is the average loss of a stochastic model selection algorithm that makes a prediction by first selecting c according to distribution Q. So we can interpret theorem 1 as a bound on the loss of a stochastic model selection algorithm using posterior Q. In the case of a countable concept class where Q is concentrated on the single concept c the quantity $D(Q||P)$ equals $\ln \frac{1}{P(c)}$ and, for large m, theorem 1 is essentially the same as lemma 1. But theorem 1 is considerably stronger than lemma 1 in that it handles the case of uncountable (continuous) concept classes. Even for countable classes theorem 1 can lead to a better guarantee than lemma 1 if the posterior Q is spread over exponentially many different models having similar empirical error rates. This might occur, for example, in mixtures of decision trees as constructed in [3, 17, 9].

The second main result of this paper is that the posterior distribution minimizing the error rate bound given in theorem 1 is a Gibbs distribution. For any value of $\beta \geq 0$ we define Q^β to be the posterior distribution defined as follows where Z is a normalizing constant.

$$dQ^\beta(c) = \frac{1}{Z} dP(c) e^{-\beta \hat{l}(c,\, S)}$$

For any posterior distribution Q define $B(Q)$ as follows.

$$B(Q) \equiv \hat{l}(Q,\, S) + \sqrt{\frac{D(Q||P) + \ln \frac{1}{\delta} + \ln m + 2}{2m-1}}$$

The second main result of the paper is the following.

Theorem 2 *If $\mathcal{C}$ is finite then there exists $\beta \geq 0$ such that Q^β is optimal, i.e., $B(Q^\beta) \leq B(Q)$ for all Q, and where β satisfies the following.*

$$\beta = 2\sqrt{(2m-1)(D(Q^\beta||P) + \ln(1/\delta) + \ln m + 2)} \tag{2}$$

Unfortunately, there can be multiple local minima in $B(Q^\beta)$ as a function of β and even multiple local minima satisfying (2). Fortunately, simpler posterior distributions achieve nearly optimal performance. To simplify the discussion we consider parameterized concept classes where each concept is specified by a parameter vector $\Theta \in R^n$. Let $l(\Theta,\ x)$ be the loss of the concept named by parameter vector Θ on the data point x (as discussed above). To further simplify the analysis we assume that for any given x we have that $l(\Theta,\ x)$ is a continuous function of Θ. For example, we might take Θ to be the coefficients of an nth order polynomial p_Θ and take $l(\Theta,\ x)$ to be $\max(1, \alpha|p_\Theta(x) - f(x)|)$ where $f(x)$ is a fixed target function and α is a fixed parameter of the loss function. Note that a two valued loss function can not be a continuous function of Θ unless the prediction is independent of Θ. Now consider a sample S consisting of m data points. These data points define an empirical loss $\hat{l}(\Theta)$ for each parameter vector Θ. This empirical loss is an average of a finite number of expressions of the form $l(\Theta,\ x)$ and hence $\hat{l}(\Theta)$ must be a continuous function of Θ. Assuming that the prior on Θ is given by a continuous density we then get that there exists a continuous density $p(\hat{l})$ on empirical errors satisfying the following where $P(U)$ denotes the measure of a subset U of the concepts according to the prior measure on concepts.

$$P(\{\Theta : \hat{l}(\Theta) \in [x,\ x+\delta]\}) = \int_x^{x+\delta} p(\hat{l})d\hat{l}$$

The second main result of the paper can be summarized as the following approximate equation where $B(Q^*)$ denotes $\inf_Q B(Q)$.

$$B(Q^*) \approx \min_{\hat{l}}\ \hat{l} + \sqrt{\frac{\ln \frac{1}{p(\hat{l})}}{2m}} \tag{3}$$

This approximate inequality is justified by the two theorems stated below. Before stating the formal theorems, however, it is interesting to compare (3) with lemma 1. For a countable concept class we can define c^* to be the concept minimizing the bound in lemma 1. For large m, lemma 1 can be interpreted as follows.

$$l(c^*) \leq \min_c \hat{l}(c, S) + \sqrt{\frac{\ln \frac{1}{P(c)}}{2m}} \tag{4}$$

Clearly there is a structural similarity between (4) and (3). However, the two formulas are fundamentally different in that (3) applies to continuous concept densities while (4) only applies to countable concept classes.

Another contribution of this paper is theorems giving upper and lower bounds on $B(Q^*)$ justifying (3). First we give a simple posterior distribution which nearly achieves the performance of (3). Define $\hat{l}^*$ as follows.

$$\hat{l}^* = \ arg \min_{\hat{l} \in [0,1]} \ \hat{l} + \sqrt{\frac{\ln \frac{1}{p(\hat{l})} + \ln \frac{1}{\delta} + \ln m + 2}{2m - 1}}$$

Define the posterior distribution $Q(\hat{l}^*)$ as follows where Z is a normalizing constant.

$$dQ(\hat{l}^*)(c) \equiv \frac{1}{Z} \begin{cases} dP(c) \ \text{ if } \ \hat{l}(c) \in [\hat{l}^*, \ \hat{l}^* + 1/m] \\ 0 \ \text{ otherwise} \end{cases}$$

We now have the following theorem.

Theorem 3 *For any prior (probability measure) on a concept class where each concept is named by a vector $\Theta \in R^n$ and any sample of m instances, if the loss function $l(\Theta, \ x)$ is always in the interval $[0, 1]$ and is continuous in Θ, the prior on Θ is a continuous probability density on R^n, $\hat{l}^* \leq 1 - 1/m$, and the density $p(\hat{l})$ is non-decreasing over the interval $[\hat{l}^*, \ \hat{l}^* + 1/m]$, then we have the following.*

$$B(Q(\hat{l}^*)) \leq \hat{l}^* + \frac{1}{m} + \sqrt{\frac{\ln \frac{1}{p(\hat{l}^*)} + \ln \frac{1}{\delta} + 2 \ln m + 2}{2m - 1}}$$

All of the assumptions used in theorem 3 are quite mild. The final assumption that the density $p(\hat{l})$ is nondecreasing over the interval defining $Q(\hat{l}^*)$ is justified by fact that the definition of $\hat{l}^*$ implies that for any differentiable density function $p(\hat{l})$ we must have that the density $p(\hat{l})$ is increasing at the point $\hat{l}^*$.

Finally we show that $Q(\hat{l}^*)$ is a nearly optimal posterior.

Theorem 4 *For any prior (probability measure) on a concept class where each concept is named by a vector $\Theta \in R^n$ and any sample of m instances, if the loss function $l(\Theta,\ x)$ is always in the interval $[0,1]$ and is continuous in Θ, and the prior on Θ is a continuous probability density on R^n, then we have the following for any posterior Q.*

$$B(Q) \geq \hat{l}^* + \sqrt{\frac{\ln \frac{1}{p(\hat{l}^*)} + \ln \frac{1}{\delta} + \ln m + 2}{2m-1}}$$

3 Related Work

A model selection guarantee very similar to (1) has been given by Barron [1]. Assume concepts $f_1,\ f_2,\ f_3,\ \ldots$ and true and empirical error rates $\epsilon(f_i)$ and $\hat{\epsilon}(f_i)$ as in (1). Let f^* be defined as follows.

$$f^* \equiv arg\min_{f_i}\ \hat{\epsilon}(f_i) + \sqrt{\frac{\ln \frac{1}{P(f_i)}}{2m}}$$

For the case of error rates (also known as 0-1 loss) Barron's theorem reduces to the following.

$$E_{S\sim D^m}\ \ \epsilon(f^*) \ \ \leq\ \ \inf_i \left(\epsilon(f_i) + \sqrt{\frac{\ln \frac{1}{P(f_i)}}{2m}}\right) + \sqrt{\frac{2\pi}{m}} \qquad (5)$$

There are several differences between (1) and (5). When discussing (1) I will take f^* to be the concept f_i minimizing the right hand side of (1) which is nearly the same as the definition of f^* in (5). Formula (1) implies the following.

$$\forall^\delta S\ \ \epsilon(f^*) \leq \inf_i \hat{\epsilon}(f_i) + \sqrt{\frac{\ln \frac{1}{P(f_i)} + \ln \frac{1}{\delta}}{2m}}$$

Note that (5) bounds the expectation of $\epsilon(f^*)$ while (1) is a large deviation result — it gives a bound on $\epsilon(f^*)$ as a function of the desired confidence level δ. Also note that (1) provides a bound on $\epsilon(f^*)$ in terms of information available in the sample while (5) provides a bound on (the expectation of) $\epsilon(f^*)$ in terms of the unknown quantities $\epsilon(f_i)$. This means that a learning algorithm based on (1) can output a performance guarantee along with the selected concept. This is true even if the concept is selected by incomplete search over the concept space and hence is different from f^*. No such guarantee can be computed from (5). If a bound in terms of the unknown quantities $\epsilon(f_i)$ is desired, the proof method used to prove (1) yields the following.

$$\forall^\delta S \;\; \epsilon(f^*) \leq \inf_i \epsilon(f_i) + \sqrt{\frac{2\left(\ln \frac{1}{P(f_i)} + \ln \frac{2}{\delta}\right)}{m}}$$

Also note that (5), like (1) but unlike theorem 1, is vacuous for continuous concept classes.

Various other model selection results similar to (1) have appeared in the literature. A guarantee involving the index of a concept in an arbitrary given sequence of concepts is given in [12]. A bound based on the index of a concept class in a sequence of classes of increasing VC dimension is given in [14]. Neither of these bounds handle an arbitrary prior distribution on concepts. They do, however, give PAC SRM performance guarantees involving some form of prior knowledge (learning bias).

Guarantees for model selection algorithms for density estimation have been given by Yamanishi [21] and Barron and Cover [2]. The guarantees bound measures of distance between a selected model distribution and the true data source distribution. In both cases the model is assumed to have been selected so as to optimize an SRM tradeoff between model complexity and the goodness of fit to the training data. The bounds hold without any assumption relating the prior distribution to the data distribution, However, the performance guarantee is better if there exist simple models that fit well. The precise statement of these bounds are somewhat involved and perhaps less interesting than the more elegant guarantee given in formula (6) discussed below.

Guarantees for model averaging have also been proved. First I will consider model averaging for density estimation. Let f_1, f_2, f_3, ... be an infinite sequence of models each of which defines a probability distribution on a set X. Let P be a "prior probability" on the densities f_i. Assume an unknown distribution g on X which need not be equal to any f_i. Let S be a sample of m elements of X sampled IID according to the distribution g. Let h be the natural "posterior" density on X defined as follows where Z is a normalizing constant.

$$h(x) \equiv \sum_i P(f_i|S) f_i(x)$$

$$P(f_i|S) \equiv \frac{1}{Z} P(f_i) P(S|f_i)$$

Note that the posterior density h is a function of the sample and hence is a random variable. Catoni [5] and Yang [23] prove somewhat different general theorems both of which have as a special case the statement that, independent of how g is selected, the expectation (over drawing a sample according to g) of the Kullback-Leibler Divergent $D(g||h)$ is bounded as follows.

$$E\ D(g||h) \leq \min_i \left(\frac{\ln \frac{1}{P(f_i)}}{m} + D(g||f_i) \right) \tag{6}$$

Again we have that (6) holds without any assumed relation between g and the prior P. If there happens to be a low complexity (simple) model f_i such that $D(g||f_i)$ is small, then the posterior density h will have small divergence from g. If no simple model has small divergence from g then $D(g||h)$ can be large. Also not that (6), unlike theorem 1, is vacuous for continuous model classes. These observations also apply to the more general forms of (6) appearing in [23] and [5]. Catoni [4] also gives performance guarantees for model averaging for density estimation over continuous model spaces using a Gibbs posterior. However, the statements of these guarantees are quite involved and the relationship to the bounds in this paper is unclear.

Yang [22] considers model averaging for prediction. Consider a fixed distribution D on pairs $\langle x,\ y\rangle$ with $x \in X$ and $y \in \{0,1\}$. Consider a countable class of conditional probability rules $f_1,\ f_2,\ f_3,\ \ldots$ where each f_i is a function from X to $[0,1]$ where $f_i(x)$ is interpreted as $P(y|x, f_i)$. Consider an arbitrary prior on the models f_i and construct the posterior given a sample S as $Q(f_i) \equiv \frac{1}{Z} P(f_i) P(S|f_i)$. This posterior on the models induces a posterior h on y given x defined as follows.

$$P(y|x, S) \equiv h(x) \equiv \sum_i Q(f_i) f_i(x)$$

Let $g(x)$ be the true conditional probability $P(y|x)$ as defined by the distribution D. For any function g' from X to $[0,1]$ define the loss $L(g')$ as follows where $x \sim D$ denotes selecting x from the marginal of D on X.

$$L(g') \equiv E_{x \sim D}\ \ |g'(x) - g(x)|^2$$

Finally, define δ_i as follows.

$$\delta_i \equiv \inf_{x \in X} \min(f_i(x),\ 1 - f_i(x))$$

For $m \geq 2$, the following is a corollary of Yang's theorem.

$$E_{S\sim D^m}\ L(h) \quad \leq 2 \inf_i \left(\frac{\ln \frac{1}{P(f_i)}}{m} + \frac{L(f_i)}{\delta_i^2} \right)$$

This formula bounds the loss of the Bayesian model average without making any assumption about the relationship between the data distributions D and the prior distribution P. However, it seems weaker than (5) or (6) in that it does not imply even for a finite model class that for large samples the loss of the posterior converges to the loss of the best model. As with (6), the guarantee is vacuous for continuous model classes. These same observations apply to the more general statement in [22].

Weighted model mixtures are also widely used in constructing algorithms with on-line guarantees. In particular, the weighted majority algorithm and its variants can be proved to compete well with the best expert on an arbitrary sequence of labeled data [13, 6, 8, 7]. The posterior weighting used in most on-line algorithms is a Gibbs posterior Q^β as defined in the statement of theorem 2. One difference between these on-line guarantees and theorem 1 is that for these algorithms one must know the appropriate value of β before seeing the training data. Since a-prior knowledge of β is required, the on-line algorithm is not guaranteed to perform well against the optimal SRM tradeoff — performing well against the optimal SRM tradeoff requires tuning β in response to the training data. Another difference between on-line guarantees and either formula (1) or theorem 1 is that (1) (or theorem 1) provides a guarantee even in cases where only incomplete searches over the concept space are feasible. On-line guarantees require that the algorithm find all concepts that perform well on the training data — finding a single simple concept that fits well is insufficient.

The most closely related earlier result is a theorem in [15] bounding the error rate of stochastic model selection in the case where the model is selected stochastically from a set U of models under a probability measure that is simply a renormalization of the prior on U. Theorem 1 is a generalization of this result to the case of arbitrary posterior distributions.

4 Proof of Theorem 1

The departure point for the proof of theorem 1 is the following where S is a sample of size m and $\Delta(c)$ abbreviates $|l(c) - \hat{l}(c, S)|$.

Lemma 2 *For any prior distribution (probability measure) P on a (possibly uncountable) concept space C we have the following.*

$$\forall^\delta S \quad E_{c\sim P}\ e^{(2m-1)\Delta(c)^2} \quad \leq \frac{4m}{\delta}$$

Proof: It suffices to prove the following.

$$E_S\ E_{c\sim P}\ e^{(2m-1)\Delta(c)^2} \leq 4m \tag{7}$$

Lemma 2 follows from (7) by an application of Markov's inequality. To prove (7) it suffices to prove the following for any individual given concept.

$$E_S\ e^{(2m-1)\Delta(c)^2} \leq 4m \tag{8}$$

For a given concept c, the probability distribution on the sample induces a probability distribution on $\Delta(c)$. By the Chernoff bound this distribution on Δ satisfies the following.

$$P(\Delta \geq x) \leq 2e^{-2mx^2} \tag{9}$$

It now suffices to show that any distribution satisfying (9) must satisfy (8). The distribution on Δ satisfying (9) and maximizing $Ee^{(2m-1)\Delta^2}$ is the continuous density $f(\Delta)$ satisfying $\int_x^\infty f(\Delta)d\Delta = 2e^{-2mx^2}$ which implies $f(\Delta) = 8m\Delta e^{-2m\Delta^2}$. So we have the following

$$\begin{aligned}
E_S\, e^{(2m-1)\Delta^2} &\leq \int_0^\infty e^{(2m-1)\Delta^2} f(\Delta)d\Delta \\
&= \int_0^\infty 8m\Delta e^{(2m-1)\Delta^2} e^{-2m\Delta^2} d\Delta \\
&= \int_0^\infty 8m\Delta e^{-\Delta^2} d\Delta \\
&= 4m
\end{aligned}$$

□

To prove theorem 1 we consider selecting a sample S. Lemma 2 implies that with probability at least $1-\delta$ over the selection of a sample S we have the following.

$$E_{c\sim P}\ e^{(2m-1)\Delta(c)^2} \leq \frac{4m}{\delta} \tag{10}$$

To prove theorem 1 it now suffices to show that the constraint (10) on the function $\Delta(c)$ implies the body of theorem 1. We are interested in computing an upper bound on the quantity $l(Q) - \hat{l}(Q,\ S)$. Note that $l(Q) - \hat{l}(Q,\ S) \leq E_{c\sim Q}|l(c_i) - \hat{l}(c_i,\ S)| = E_{c\sim Q}\Delta(c)$. We now prove the following lemma.

Lemma 3 *For $\beta > 0$, $K > 0$, and $Q, P, \Delta \in R^n$ satisfying $P_i \geq 0$, $Q_i \geq 0$, $\Delta_i \geq 0$, and $\sum_{i=1}^n Q^i = 1$, we have that if*

$$\sum_{i=1}^n P_i e^{\beta \Delta_i^2} \leq K$$

then

$$\sum_{i=1}^n Q_i \Delta_i \leq \sqrt{\frac{D(Q||P) + \ln K}{\beta}}.$$

Before proving lemma 3 we note that lemmas 3 and 2 together imply theorem 1. To see this consider a sample satisfying (10) and an arbitrary posterior probability measure Q on concepts. It is possible to define three infinite sequences of vectors Q^1, Q^2, Q^3, ..., P^1, P^2, P^3 ..., and Δ^1, Δ^2, Δ^3, ... such that Q^n, P^n, and Δ^n satisfy the conditions of lemma 3 with $K = 4m/\delta$ and $\beta = 2m - 1$ and satisfying the following.

$$E_{c\sim Q}\Delta(c) = \lim_{n\to\infty} \sum_{i=1}^n Q_i^n \Delta_i^n$$

$$D(Q||P) = \lim_{n\to\infty} \sum_{i=1}^n Q_i^n \ln \frac{Q_i^n}{P_i^n}$$

By taking the limit of the conclusion of lemma 3 we then get $E_{c\sim Q}\Delta(c) \leq \sqrt{(D(Q||P) + \ln(1/\delta) + \ln m + 2)/(2m-1)}$.

To prove lemma 3 it suffices to consider only those values of i for which $Q_i > 0$ — dropping the indices where $Q_i = 0$ does not change the value of $\sum_{i=1}^n Q_i \Delta_i$ while enlarging the feasible set by weakening the constraint (10). Furthermore, if $P_i = 0$ at some point where $Q_i > 0$ then $D(Q||P) = \infty$ and the theorem is immediate. So we can assume without loss of generality that $Q_i > 0$ and $P_i > 0$ for all i.

By Jensen's inequality we have $(\sum_{i=1}^n Q_i \Delta_i)^2 \leq \sum_{i=1}^n Q_i \Delta_i^2$. So it now suffices to prove that $\sum_{i=1}^n Q_i \Delta_i^2 \leq (D(Q||P) + \ln K)/\beta$. This is a consequence of the following lemma.[1]

[1]The original version of this paper [16] proved a bound of approximately the form $\hat{l}(Q) + \sum_{i=1}^n Q_i \sqrt{\ln(Q_i/P_i)/(2m)}$ by maximizing $\sum_{i=1}^n Q_i \Delta_i$ subject to constraint 10. A

Lemma 4 *For $\beta > 0$, $K > 0$, and $Q, P, y \in R^n$ satisfying $P_i > 0$, $Q_i > 0$, and $\sum_{i=1}^n Q_i = 1$, if*

$$\sum_{i=1}^{n} P_i e^{\beta y_i} \leq K \tag{11}$$

then

$$\sum_{i=1}^{n} Q_i y_i \leq \frac{D(Q||P) + \ln K}{\beta}$$

To prove lemma 4 we take P and Q as given and use the Kuhn-Tucker conditions to find a vector y maximizing $\sum_{i=1}^n Q_i y_i$ subject to the constraint (11).

Lemma 5 (Kuhn-Tucker) *If C and $f_1, \ldots, f_n$ are functions from R^n to R, y is a maximum of $C(y)$ over the set satisfying $f_1(y) \leq 0, \ldots, f_n(y) \leq 0$, and C and each f_i are continuous and differentiable at y, then either $\nabla C = 0$ (at y), or there exists some f_i with $f_i(y) = 0$ and $\nabla f_i = 0$ (at y), or there exists a nonempty subset of the constraints $f_{i_1}(y) \leq 0, \ldots, f_{i_k}(0) \leq 0$ such that $f_{i_j}(y) = 0$ for $1 \leq j \leq k$ and positive coefficients $\lambda_1, \ldots, \lambda_k$ such that $\nabla C = \lambda_1 \nabla f_{i_1} + \ldots + \lambda_n \nabla f_{i_n}$ (at y).*

Note that lemma 4 allows y_i to be negative. The first step in proving lemma 4 is to show that without loss of generality we can work with a closed and compact feasible set. For $K > 0$ it is not difficult to show that there exists a feasible point, i.e., a vector y such that $\sum_{i=1}^n P_i e^{\beta y_i} \leq K$. Let C_0 denote an arbitrary feasible value, i.e., $\sum_{i=1}^n Q_i y_i$ for some feasible point y. Without loss of generality we need only consider points y satisfying $\sum_{i=1}^n Q_i y_i \geq C_0 - 1$. So we now have a constrained optimization problem with objective function $\sum_{i=1}^n Q_i y_i$ and feasible set defined by the following constraints.

$$\sum_{i=1}^{n} P_i e^{\beta y_i} \leq K \tag{12}$$

$$\sum_{i=1}^{n} Q_i y_i \geq C_0 - 1 \tag{13}$$

---

version of theorem 1, which is of the form $\hat{l}(Q, S) + \sqrt{(\sum_{i=1}^n Q_i \ln(Q_i/P_i))/(2m)}$, was then proved from this bound by an application of Jensen's inequality. The idea of maximizing $\sum_{i=1}^n Q_i \Delta_i^2$ and achieving theorem 1 directly is due to Robert Schapire.

Constraint (12) implies an upper bound on each y_i and constraint (13) then implies a lower bound on each y_i. Hence the feasible set is closed and compact.

We now note that any continuous objective function on a closed and compact feasible set must be bounded and must achieve its maximum value on some point in the set. A constraint of the form $f(y) \leq 0$ will be called active at y if $f(y) = 0$. For an objective function whose gradient is nonzero everywhere, at least one constraint must be active at the maximum. Since C_0 is a feasible value of the objective function, constraint (13) can not be active at the maximum. So by the Kuhn-Tucker lemma, the point y achieving the maximum value must satisfy the following.

$$Q_i = \lambda P_i \beta e^{\beta y_i}$$

Which implies the following.

$$y_i = \frac{\ln(\frac{Q_i}{\lambda P_i \beta})}{\beta}$$

Since constraint (12) must be active at the maximum, we have the following.

$$\sum_{i=1}^{n} P_i e^{\beta y_i} = \sum_{i=1}^{n} \frac{Q_i}{\lambda \beta} = \frac{1}{\lambda \beta} = K$$

So we get $\lambda = 1/(\beta K)$ and the following.

$$\sum_{i=1}^{n} Q_i y_i = \sum_{i=1}^{n} Q_i \frac{\ln \frac{Q_i}{P_i} + \ln K}{\beta} = \frac{D(Q||P) + \ln K}{\beta}$$

Since this is the maximum value of $\sum_{i=1}^{n} Q_i y_i$, the lemma is proved.

5 Proof of Theorem 2

We wish to find a distribution Q minimizing $B(Q)$ defined as follows where the distribution P and the empirical error $\hat{l}(c)$ are given and fixed.

$$B(Q) \equiv E_{c \sim Q} \;\; \hat{l}(c) + \sqrt{\frac{D(Q||P) + \ln \frac{1}{\delta} + \ln m + 2}{2m - 1}}$$

Letting K be $\ln(1/\delta)+\ln m+2$ and letting γ be $2m-1$ this objective function can be rewritten as follows where K and γ are fixed positive quantities independent of Q.

$$B(Q) = E_{c\sim Q}\hat{l}(c) + \sqrt{\frac{D(Q||P)+K}{\gamma}}$$

To simplify the analysis we consider only finite concept classes. Let P_i be the prior probability of the ith concept and let $\hat{l}_i$ be the empirical error rate of the ith concept. The problem now becomes finding values of Q_i satisfying $Q_i \geq 0$ and $\sum_i Q_i = 1$ minimizing the following.

$$B(Q) = \sum_i Q_i \hat{l}_i + \sqrt{\frac{D(Q||P)+K}{\gamma}}$$

If P_i is zero then if Q_i is nonzero we have that $D(Q||P)$ is infinite. So for minimizing $B(Q)$ we can assume that Q_i is zero if P_i is zero and we can assume without loss of generality that all P_i are nonzero. If all P_i are nonzero then the objective function is a continuous function of a compact feasible set and hence realizes its minimum at some point in the feasible set. Now consider the following partial derivative.

$$\begin{aligned}
\frac{\partial D(Q||P)}{\partial Q_i} &= \frac{\partial\ \sum_j Q_j \ln \frac{Q_j}{P_j}}{\partial Q_i} \\
&= \frac{\partial\ Q_i \ln \frac{Q_i}{P_i}}{\partial Q_i} \\
&= 1 + \ln \frac{Q_i}{P_i}
\end{aligned}$$

Note that if Q_i is zero when P_i is nonzero then $\partial D(Q||P)/\partial Q_i = -\infty$. This means that any transfer of an infinitesimal quantity of probability mass to Q_i reduces the bound. So the minimum must not occur at a boundary point satisfying $Q_i = 0$. So we can assume without loss of generality that Q_i is nonzero for each i where P_i is nonzero — the two distributions have the same support. The Kuhn-Tucker conditions then imply that $\nabla B = 0$ or ∇B is in the direction of the gradient of one of the constraints $\sum_i Q_i \leq 1$ or

$\sum_i Q_i \geq 1$. In all of these cases there must exist a single value λ such that for all i we have $\partial B/\partial Q_i = \lambda$. This yields the following.

$$\begin{aligned}
\lambda &= \frac{\partial B}{\partial Q_i} \\
&= \hat{l}_i + \frac{1}{2}\left(\frac{D(Q||P)+K}{\gamma}\right)^{-1/2}\frac{1}{\gamma}\frac{\partial D(Q||P)}{\partial Q_i} \\
&= \hat{l}_i + \frac{1}{2}\left(\frac{D(Q||P)+K}{\gamma}\right)^{-1/2}\frac{1}{\gamma}\left(\ln\frac{Q_i}{P_i}+1\right) \\
\ln\frac{Q_i}{P_i} &= (\lambda - \hat{l}_i)2\sqrt{\gamma(D(Q||P)+K)} - 1 \\
Q_i &= P_i \exp\left((\lambda - \hat{l}_i)2\sqrt{\gamma(D(Q||P)+K)} - 1\right)
\end{aligned}$$

Hence the minimal distribution has the following form.

$$Q_i = \frac{1}{Z}P_i e^{-\beta \hat{l}_i}$$

$$\beta = 2\sqrt{\gamma(D(Q||P)+k)} = 2\sqrt{(2m-1)(D(Q||P)+\ln\frac{1}{\delta}+\ln m + 2)}$$

This is the distribution Q^β of theorem 2.

6 Proof of Theorems 3 and 4

Let $Q(\hat{l}^*)$ be the posterior distribution of theorem 3. First we note the following.

$$\begin{aligned}
D(Q(\hat{l}^*)||P) &= E_{c\sim Q(\hat{l}^*)} \ln\frac{dQ(c)}{dP(c)} \\
&= E_{c\sim Q(\hat{l}^*)} \ln\frac{1}{Z} \\
&= \ln\frac{1}{P(\{c:\ \hat{l}(c)\in[\hat{l}^*,\ \hat{l}^*+1/m]\})}
\end{aligned}$$

We have assumed that $p(\hat{l})$ is nondecreasing over the interval $[\hat{l}^*,\ \hat{l}^*+1/m]$. This implies the following.

$$P(\{c:\ \hat{l}(c)\in[\hat{l}^*,\ \hat{l}^*+1/m]\}] \geq \frac{1}{m}p(\hat{l}^*)$$

$$D(Q(\hat{l}^*)||P) \leq \ln \frac{1}{p(\hat{l}^*)} + \ln m$$

We also have that $\hat{l}(Q(\hat{l}^*)) \leq \hat{l}^* + 1/m$ and theorem 3 now follows from the definition of $B(Q)$.

We now prove theorem 4. First we define a concept distribution U such that U induces a uniform distribution on those error rates $\hat{l}$ with $p(\hat{l}) > 0$. Let W be the subset of the values $\hat{l} \in [0,1]$ such that $p(\hat{l}) > 0$. Let α denote the size of W as measured by the uniform measure on $[0, 1]$. Note that $\alpha \leq 1$. Define the concept distribution U as follows.

$$dU(c) = dP(c) \begin{cases} \frac{1}{\alpha p(\hat{l}(c,\ S))} & \text{if } p(\hat{l}(c,\ S)) > 0 \\ 0 & \text{otherwise} \end{cases}$$

The total measure of U can be written as follows.

$$\int_{\hat{l} \in W} \frac{dU}{dP} \frac{dP}{d\hat{l}} d\hat{l} = \int_{\hat{l} \in W} \frac{1}{\alpha p(\hat{l})} p(\hat{l}) d\hat{l} = 1$$

Hence U is a probability measure on concepts.

Now let Q be an arbitrary posterior distribution on concepts. We have the following.

$$\begin{aligned} D(Q||P) &= E_{c \sim Q} \ln \frac{dQ}{dP} \\ &= E_{c \sim Q} \left[\ln \frac{dU}{dP} + \ln \frac{dQ}{dU} \right] \\ &= E_{c \sim Q} \ln(1/p(\hat{l}(c,\ S))) + \ln(1/\alpha) + D(Q||U) \\ &\geq E_{c \sim Q} \ln(1/p(\hat{l}(c,\ S))) \end{aligned}$$

This implies the following where the third line follows from Jensen's inequality.

$$\begin{aligned} B(Q) &= \hat{l}(Q,\ S) + \sqrt{\frac{D(Q||P) + \ln \frac{1}{\delta} + \ln m + 2}{2m - 1}} \\ &\geq \left[E_{c \sim Q}\ \ \hat{l}(c,\ S) \right] + \sqrt{\frac{E_{c \sim Q} \ln(1/p(\hat{l}(c,\ S))) + \ln \frac{1}{\delta} + \ln m + 2}{2m - 1}} \end{aligned}$$

$$
\begin{aligned}
&\geq \quad E_{c\sim Q}\left[\hat{l}(c,\,S)+\sqrt{\frac{\ln(1/p(\hat{l}(c,\,S)))+\ln\frac{1}{\delta}+\ln m+2}{2m-1}}\right] \\
&\geq \quad \min_{\hat{l}}\quad \hat{l}+\sqrt{\frac{\ln(1/p(\hat{l}))+\ln\frac{1}{\delta}+\ln m+2}{2m-1}}
\end{aligned}
$$

7 Conclusion

PAC-Bayesian learning algorithms combine the flexibility prior distribution on models with the performance guarantees of PAC algorithms. PAC-Bayesian Stochastic model selection can be given performance guarantees superior to analogous guarantees for deterministic PAC-Bayesian model selection. The performance guarantees for stochastic model selection naturally handle continuous concept classes and lead to a natural notion of an optimal posterior distribution to use in stochastically selecting a model. Although the optimal posterior is a Gibbs distribution, it is shown that under mild assumptions simpler posterior distributions perform nearly as well. An open question is whether better guarantees can be given for model averaging rather than stochastic model selection.

Acknowledgments: I would like to give special thanks to Manfred Warmuth for inspiring this paper and emphasizing the analogy between the PAC and on-line settings. I would also like to give special thanks to Robert Schapire for simplifying and strengthening theorem 1. Avrim Blum, Yoav Freund, Michael Kearns, John Langford, Yishay Mansour, and Yoram Singer also provided useful comments and suggestions.

References

[1] A.R. Barron. Complexity regularization with application to artificial neural networks. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 561–576. Kluwer Academic Publishers, 1991.

[2] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.

[3] Wray Buntine. Learning classification trees. *Statistics and Computing*, 2:63–73, 1992.

[4] Olivier Catoni. Gibbs estimators. to appear in Probability Theory and Related Fields.

[5] Olivier Catoni. Universal aggregation rules with sharp oracle inequalities. to appear in Annals of Statistics.

[6] Nicol Cesa-Bianchi, Yoav Freund, David P. Helmbold, David Haussler, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *JACM*, 44(3):427–485, 1997.

[7] Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, pages 79–103, 1999.

[8] Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. Using and combining predictors that specialize. In *COLT-97*, 1997.

[9] David P. Helmbold and Robert E. Schapire. Predicting nearly as well as the best pruning of a decision tree. *Machine Learning*, 27(1):51–68, 1997.

[10] Michael Kearns and Yishai Mansour. A fast, bottom-up decision tree pruning algorithm with near-optimal generalization. In *Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann, 1998.

[11] Michael Kearns, Yishai Mansour, Andrew Ng, and Dana Ron. An experimental and theoretical comparison of model selection methods. In *Proceedings of the Eighth ACM Conference on Computational Learning Theory*, pages 21–30. ACM Press, 1995.

[12] Nathan Linial, Yishay Mansour, and Ronald Rivest. Results on learnability and the Vapnik-Chervonenkis dimension. *Information and Computation*, 90:33–49, 1991.

[13] N. Littlestone and Manfred Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994. An extended abstract appeared in COLT-89.

[14] G. Lugosi and K. Zeger. Concept learning using complexity regularization. *IEEE Transactions on Information Theory*, 42:48–54, 1996.

[15] David McAllester. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, 1998.

[16] David McAllester. Pac-bayesian model averaging. In *COLT-99*, 1999.

[17] Jonathan J. Oliver and David Hand. On pruning and averaging decision trees. In *Proceedings of the Twelfth International Conference on Machine Learning*, 1995.

[18] F.C. Pereira and Y. Singer. An efficient extension to mixture techniques for prediction and decision trees. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 114–121, 1997. Long version to appear in Machine Learning.

[19] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[20] John Shawe-Taylor and Robert Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*. ACM Press, 1997.

[21] Kenji Yamanishi. Learning non-parametric densities in tyerms of finite-dimensional parametric hypotheses. *IEICE Trans. Inf. and Syst.*, E75-D(4), 1992.

[22] Yuhang Yang. Adaptive estimation in pattern recognition by combining different procedures. *Statistica Sinica*, 10:1069–1089, 2000.

[23] Yuhang Yang. Mixing strategies for density estimation. *Annals of Statistics*, 28:75–87, 2000.